DOCUMENT RESUME

ED 066 297 SE 014 177

AUTHOR	Egelston, Richard L.; Egelston, Judy C.
TITLE	Self-Evaluation and Performance on Classroom Tests.
PUB DATE	Apr 72
NOTE	14p.; Paper presented at the National Association for Research in Science Teaching meeting, Chicago, Illinois, April 1972
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Academic Achievement; *Evaluation; *Grade Prediction; Secondary School Science; *Self Evaluation; Testing

ABSTRACT

In an investigation of the accuracy of self-evaluation on test performance, 210 junior high school science students were asked to predict their scores before and after taking each unit test. Absolute differences between the two predictions and actual scores were the random variables analyzed. Analysis of variance and Markov chain analyses revealed significant differences by achievement level, practice, and in rate of learned and perhaps should be incorporated into the school curriculum. (Author/CP)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED 066297

SELF-EVALUATION AND PERFORMANCE ON CLASSROOM TESTS

Richard L. Egelston and Judy C. Egelston

State University of New York at Geneseo

When people leave the formal educational setting and enter the worlds of work and leisure, they are required to make many decisions based upon their own abilities and interests. Each of the decisions requires some assessment about the degree of success or enjoyment in the activity in which they are to become engaged. Hopefully, the evaluation of the potential activity will be rational and based upon a thorough knowledge of personal capabilities. However, self-evaluation processes may be difficult to learn and may need to be developed and taught within the school curriculum.

Research on self-evaluation is meager, and that which has been done generally involves simple tasks not at all comparable to the complex activities which individuals undertake in later life. Such studies have been typified by tasks involving the persuit roter (Rotter, 1942) and number cancellation tasks (Anderson and Brandt, 1939). While it is possible to construct good experimental controls with these simple tasks, the meaningfulness of the tasks for the subjects is somewhat questionable, and any inferences drawn from these studies toward level of aspiration or self-evaluation are highly suspect. One meaningful task in the school setting which is repetitive enough for studying self-evaluation is that of test taking.

Murstein (1965) found that neither high nor low achieving college

students changed their predictions of final grades as a result of feedback on mid-semester examinations. This result was not confirmed by Wolfe (in press) who found that college students became more accurate predictors as a result of mid-semester feedback.

In an attempt to determine the influence of sex and achievement on the ability to predict test scores for college students, Sumner and Johnson (1949) found discrepancy scores to be less for high achieving students than for low achieving students. They also found that females of all quartile levels are more accurate predictors than males of comparable levels.

With secondary school students, Pickup and Anthony (1968) found that females who predicted higher scores than they received tended to reduce subsequent predictions while males did not. Low achievers were more likely to predict higher scores than they received than did high achievers.

Pennington's (1940) experiments on college students indicated that failure resulted in a lower level of aspiration, and success (passing with high grades) resulted in an upward swing in predicted scores on the following examination. With fifth grade children, Anderson and Brandt (1939) found that poor students set goals consistently above past performance, and good students set goals consistently below past performance.

Utilizing the concepts involved in self-evaluation is a task of the problem solving order as described by Gagné (1965), and involves a great deal of formal reasoning. Inhelder and Piaget (1958) have found that formal reasoning procedures typically begin at age 11 or 12 and build

up to a plateau at about age 14 or 15. Since students of this age are normally found in the junior high school, maturational differences were expected.

Several hypotheses were examined in this study: Whether or not students in differing achievement quartiles were able to self-evaluate more accurately; whether experiencing the task of taking the test made any difference in the ability to self-evaluate; whether students in the differing quartiles would improve more and at differing rates with practice, and whether sex made any difference in the ability to self-evaluate.

Accurate self-evaluation of pretest performance required the subject to recognize how much information he understood in comparison with what he thought the teacher expected him to know. Few cues were available except for the style and quantity of class review prior to the test, and the practice of making predictions. Additional cues were available for the posttest predictions such as the number, difficulty and style of the items as well as the practice effects. If subjects attended to the cues, it was expected that their accuracy would increase from pretest to posttest prediction. Also, if the students attended to the cues, a practice effect would probably be demonstrated.

Method

Two hundred ten students in eight general science classes and one earth science class from a rural Eastern New York secondary school were used as subjects. All students were in grades 7-9. Classes varied in size from sixteen to thirty-two students and were taught by two teachers. Within each grade the top one-fourth of the students were homogeneously

grouped for enrichment courses and the remaining students were divided into two sections of comparable ability.

At the beginning of the school year the teachers explained to the students that on each unit test the students would be asked to predict the percentage score they would get on the test immediately before (pretest prediction) and immediately after (posttest prediction) taking the test. Separate slips of paper were stapled to the test for the pretest guess, and when filled out were torn off and collected. Space was available on the test booklets for recording the posttest predictions. Students were told to base their predictions upon how well they understood the material and how difficult they thought the test would be (or was). Reminders were frequently given that the predictions would not effect actual grades in any way. Care was taken not to provide feedback on the accuracy of prediction, although test results were returned as soon as possible.

Absolute differences between each predicted score and the actual score for the test were used as random variables.

The number of tests given to each class ranged between eight and thirteen. All tests were constructed to be somewhat discriminatory in nature, and perfect scores were rarely achieved.

In the few cases where a subject failed to make a prediction, the mean prediction was used and was derived from all the pretest or posttest predicted scores the subject did make.

Within each section subjects were ranked from high to low on the final examination. Each section was then divided into four achievement levels called quartiles. Within each section, however, the quartiles

were unequal in size due to tied scores and the total section size not being divisible by four.

For each of the nine sections a three way nonorthogonal trend analysis of variance was conducted. Factor A was the quartile level of the subjects, factor B was the pretest and posttest (time) prediction, and factor C (the trend factor) was the sequence of tests taken.

Tests of hypotheses were performed in the following order: (a) A x C linear, quadratic and cubic trend interactions, (b) C linear, quadratic and cubic trends, and (c) A, B, and the C_k-C_1 contrast. The first hypothesis was tested in all six arrangements with the other hypotheses placed in a particular order. Whether or not significant interactions were present, tests of the main effects were made in all possible orders. In no case were the residual trend components or the residual trend interaction components tested for significance. The assumption was made that each successive practice trial was equally effective in producing an increment in the ability to self-evaluate, although the time intervals between tests were unequal.

All hypotheses were tested at the five percent level of significance.

According to Rotter (1942) and others, predicted scores are often dependent upon the actual performance of the previous trial. Since achievement scores are somewhat related from test to test, it is not unreasonable that predictions will be related to one another, and that discrepancy scores will be mediated by both achievement and previous predictions. The assumption was made that the discrepancy score for trial t+1 was conditional upon the discrepancy score for trial t, for a second analysis of pretest and posttest predictions.

A vector of discrepancy scores was constructed for each student and the data coded as conditional frequencies with a five point interval. The data for all students in each section were pooled and conditional probability matrices (transition matrices) were derived. A Markov chain analysis provided limiting vectors of probabilities (tolerance = .0005) for each section. (The limiting vector provides an estimate of the proportion of time the group will predict any category over an infinite number of trials.) The limiting vectors were converted to cumulative probability vectors and the pretest vector was compared with the posttest vector via a Kolmogorov-Smirnov Two Sample Test.

Results

Significant differences were found among the quartiles (A) within seven of the nine sections and between the two times of prediction (B) for three sections. No significant A x B interactions were found (see Table 1). Apparently students of differing achievement levels within

---

Insert Table 1 about here

---

the same section are not equal in the ability to self-evaluate. Generally the higher achieving students were more accurate than the lower achieving students. For many students, taking the test did not allow for a more accurate self-appraisal (before feedback) than before taking the test. Furthermore, the improvement from pretest to posttest prediction remains relatively constant for all ability students. Table 2 summarizes the trend analyses for the nine sections.

Insert Table 2 about here

The differences in trend components are relatively unimportant and may be explained by two factors: differences in degree of self-assurance in understanding the various units required, and the differential difficulty of the tests.

Four of the nine sections displayed significant quartile by test interactions indicating differing rates of improvement following practice. Each of the four sections was composed of heterogeneously grouped students and contained a larger range of ability than the homogeneously sectioned students. If the sections had been chosen without regard to ability, it is likely that more sections would have produced significant interactions. It might well be that differences in ability need to be quite large before differences in the rate of improvement will be demonstrated within a classroom.

Within the same trend analyses, contrasts of the last predictions with the first predictions were conducted, and found to be more accurate at the end of the year in seven of nine sections. Thus, practice tends to improve accuracy of self-evaluation.

Two way analyses of variance (sex by time of prediction) were performed after pooling data across tests and quartiles within each section. In no instance was a significant difference found between males and females.

For the additional pretest-posttest analysis the cumulative proportion vectors derived from the Markov chain analysis are illustrated

in Table 3 for each section. In sections 7, 8, and 9 (all of grade 9) the posttest predictions were significantly more accurate than the pretest predictions. Although the data were pooled over quartiles and

Insert Table 3 about here

tests, it would appear that grade 9 students attend more to the cues necessary for comparing their knowledge with what is called for in the test questions. It should be pointed out again that the proportions given in Table 3 are long range estimates of performance. It may be that ninth graders are more conscious of the importance of school work than seventh or eighth graders, or it may be that a higher level of intellectual maturity is necessary as Inhelder and Piaget (1958) have suggested.

Conclusions

The findings of this study are suggestive rather than definitive and generalization to the population of junior high students is perilous. Nevertheless it appears that some students in these grades can learn to improve their evaluations of self-performance on cognitive tasks.

The results of this study suggest that high achieving students are more accurate at self-evaluation and that they improve at a faster rate than low achieving students. Practice tends to improve the accuracy of prediction and under some conditions *a' posteriori* assessment may be more accurate than *a' priori* assessment. The two sexes do not appear to be different in their ability to assess their own performance.

With the great emphasis on rational decision making, it would seem important to examine personal capabilities and personal performance in an objective light. Therefore, accurate self-evaluation appears to be a reasonable process to incorporate into the school curriculum. Science classes may be the logical place to undertake this instruction, since objective measurement forms one of the cornerstones of this field.

TABLE 1

Summary Of Non-orthogonal Trend Analyses For Each Section (Part A)*

Section	N Tests	$\bar{Q}_1$ n_1	$\bar{Q}_2$ n_2	$\bar{Q}_3$ n_3	$\bar{Q}_4$ n_4	Levels ~~Quartiles~~ (A)	Pre-Post (B)	A x B
1	8	5	4	5	5	<.05	ns	ns
2	10	4	5	6	3	<.05	ns	ns
3	9	5	4	3	4	ns	ns	ns
4	8	6	6	7	5	<.05	ns	ns
5	8	7	8	6	7	<.05	<.05	ns
6	11	4	4	5	4	ns	ns	ns
7	8	8	7	7	8	<.05	<.05	ns
8	12	8	8	10	6	<.05	ns	ns
9	13	6	7	7	6	<.05	<.05	ns

*A = Quartiles factor, B = Time of prediction factor.

TABLE 2

Summary Of Non-Orthogonal Trend Analyses For Each Section (Part B)*

Section	A x C Cub Int	A x C Quad Int	A x C Lin Int	C Cub	C Quad	C Lin	First test Last test Contrast
1	ns	ns	ns	<.05	<.05	ns	<.05
2	ns	<.05	ns	<.05	<.05	<.05	<.05
3	ns	ns	ns	ns	ns	ns	<.05
4	ns	ns	<.05	ns	ns	ns	<.05
5	ns	ns	ns	ns	ns	ns	ns
6	ns	ns	ns	ns	ns	<.05	ns
7	ns	ns	<.05	ns	ns	ns	<.05
8	ns	ns	<.05	<.05	ns	<.05	<.05
9	ns	ns	ns	<.05	ns	<.05	<.05

*A = Quartiles factor, C = Tests factor

TABLE 3

Cumulative Proportion Vectors For Kolmogorov-Smirnov Two Sample Tests By Section

		Discrepancies in percentage points						
	Section	0-5	6-10	11-15	16-20	21-25	over 25	χ^2
1	Pretest	.27	.43	.55	.68	.79	1.00	.62
	Posttest	.25	.46	.60	.72	.80	1.00	
2	Pretest	.36	.64	.72	.85	.96	1.00	.79
	Posttest	.35	.61	.77	.85	.93	1.00	
3	Pretest	.31	.59	.84	.92	.96	1.00	3.10
	Posttest	.42	.69	.83	.92	.96	1.00	
4	Pretest	.23	.48	.61	.74	.80	1.00	.83
	Posttest	.27	.50	.65	.75	.85	1.00	
5	Pretest	.26	.47	.63	.73	.83	1.00	3.11
	Posttest	.32	.55	.72	.80	.90	1.00	
6	Pretest	.26	.45	.62	.74	.82	1.00	.13
	Posttest	.26	.46	.60	.74	.82	1.00	
7	Pretest	.22	.40	.55	.68	.79	1.00	11.47*
	Posttest	.34	.56	.72	.85	.93	1.00	
8	Pretest	.26	.42	.54	.74	.83	1.00	10.08*
	Posttest	.29	.52	.66	.80	.86	1.00	
9	Pretest	.34	.61	.75	.85	.90	1.00	6.13*
	Posttest	.44	.67	.83	.90	.95	1.00	

*$p < .05$ with 2 df.

REFERENCES

Anderson, H. H. and Brandt, H. F., "A Study of Motivation, Involving Self-Announced Goals of Fifth-Grade Children and the Concept of Level of Aspiration," Journal of Social Psychology, X (1939), pp. 209-232.

Gagné, R. M., Conditions of Learning. New York: Holt, Rinehart and Winston, 1965.

Inhelder, B. and Piaget, J., The Growth of Logical Thinking. New York: Basic Books, 1968, p. 347.

Murstein, B., "The Relationship of Grade Expectations and Grades Believed to be Deserved to Actual Grades Received," Journal of Experimental Education, XXXIII (1965), pp. 357-362.

Pennington, L. A., "Shifts in Aspiration Level After Success and Failure in the College Classroom," Journal of General Psychology, XXXIII (1940), pp. 305-313.

Pickup, A. J., and Anthony W. S., "Teachers' Marks and Pupil's Expectations: The Short-term Effects of Discrepancies Upon Classroom Performance in Secondary Schools," British Journal of Educational Psychology, XXXVIII (1968), pp. 302-309.

Rosenfeld, H. and Zander, A., "The Influence of Teachers on Aspirations of Students," Journal of Educational Psychology, LII (1961), pp. 1-11.

Rotter, J. B., "Level of Aspiration as a Method of Studying Personality, I. A Critical Review of Methodology," Psychological Review, XLIX (1942), pp. 463-474.

Sumner, F. C., and Johnson, E. C., "Sex Differences in Levels of Aspiration and in Self-estimates of Performance in a Classroom Situation," Journal of Psychology, XXVII (1949), pp. 483-490.

Wolfe, R. B., "Perceived Locus of Control and Prediction of Own Academic Performance," Journal of Consulting and Clinical Psychology, (in press).